AFRL-RX-WP-TP-2010-4163

OPTIMAL ALLOCATION OF TESTING RESOURCES FOR STATISTICAL SIMULATIONS (PREPRINT)

Patrick Golden

Metals Branch
Metals, Ceramics & NDE Division

Carolina Quintana, Gulshan Singh, and Harry R. Millwater

University of Texas at San Antonio

APRIL 2010

Approved for public release; distribution unlimited.

See additional restrictions described on inside pages

STINFO COPY

AIR FORCE RESEARCH LABORATORY
MATERIALS AND MANUFACTURING DIRECTORATE
WRIGHT-PATTERSON AIR FORCE BASE, OH 45433-7750
AIR FORCE MATERIEL COMMAND
UNITED STATES AIR FORCE

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
April 2010	Journal Article Preprint	01 April 2010 – 30 April 2010

4. TITLE AND SUBTITLE
OPTIMAL ALLOCATION OF TESTING RESOURCES FOR STATISTICAL SIMULATIONS (PREPRINT)

5a. CONTRACT NUMBER
In-house

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
62102F

6. AUTHOR(S)
Patrick Golden (AFRL/RXLMN)
Carolina Quintana, Gulshan Singh, and Harry R. Millwater (University of Texas at San Antonio)

5d. PROJECT NUMBER
4347

5e. TASK NUMBER
RG

5f. WORK UNIT NUMBER
M02R1000

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Metals Branch (AFRL/RXLM)
Metals, Ceramics & NDE Division
Materials and Manufacturing Directorate
Wright-Patterson Air Force Base, OH 45433-7750
Air Force Materiel Command, United States Air Force

University of Texas at San Antonio
San Antonio, TX 78249

8. PERFORMING ORGANIZATION REPORT NUMBER
AFRL-RX-WP-TP-2010-4163

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Air Force Research Laboratory
Materials and Manufacturing Directorate
Wright-Patterson Air Force Base, OH 45433-7750
Air Force Materiel Command
United States Air Force

10. SPONSORING/MONITORING AGENCY ACRONYM(S)
AFRL/RXLMN

11. SPONSORING/MONITORING AGENCY REPORT NUMBER(S)
AFRL-RX-WP-TP-2010-4163

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited.

13. SUPPLEMENTARY NOTES
Journal article submitted to *AIAA Journal*. PAO Case Number: 88ABW-2010-1272; Clearance Date: 17 Mar 2010. Paper contains color.

The U.S. Government is joint author of this work and has the right to use, modify, reproduce, release, perform, display, or disclose the work.

14. ABSTRACT
It is well known that statistical estimates from simulation, e.g., computation of mean and standard deviation, often involve significant uncertainty due to uncertainty in the input parameters. The uncertainty in the output results can be reduced with additional data on the inputs, e.g., additional experiments. An optimization methodology is proposed and implemented to determine the optimal additional experiments to conduct in order to minimize the variance in the output mean and standard deviation subject to a cost constraint. The number of additional experiments to add for each random variable depends upon several factors: the number of initial data points, the importance of the random variable in the response, the range of the random variable, and the cost of each experiment. The methodology is demonstrated using several numerical examples. The results indicate that the particle swarm optimization performs well and the solutions obtained are superior to ad hoc ones.

15. SUBJECT TERMS
statistical estimates, optimization methodology

16. SECURITY CLASSIFICATION OF: a. REPORT	b. ABSTRACT	c. THIS PAGE	17. LIMITATION OF ABSTRACT:	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON (Monitor)
Unclassified	Unclassified	Unclassified	SAR	40	Patrick Golden
					19b. TELEPHONE NUMBER *(Include Area Code)* N/A

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39-18

OPTIMAL ALLOCATION OF TESTING RESOURCES FOR STATISTICAL SIMULATIONS

Carolina Quintana[1], Gulshan Singh[2], and Harry R. Millwater[3]
University of Texas at San Antonio, TX 78249, USA

Patrick Golden[4]
Materials and Manufacturing Directorate, Air Force Research Laboratory, Wright-Patterson AFB, OH 45433, USA

It is well known that statistical estimates from simulation, e.g. computation of mean and standard deviation, often involve significant uncertainty due to uncertainty in the input parameters. The uncertainty in the output results can be reduced with additional data on the inputs, e.g., additional experiments. An optimization methodology is proposed and implemented to determine the optimal additional experiments to conduct in order to minimize the variance in the output mean and standard deviation subject to a cost constraint. The number of additional experiments to add for each random variable depends upon several factors: the number of initial data points, the importance of the random variable in the response, the range of the random variable, and the cost of each experiment. The methodology is demonstrated using several numerical examples. The results indicate that the particle swarm optimization performs well and the solutions obtained are superior to ad hoc ones.

1. Nomenclature

A_i	=	Constants in the response function for variable X_i
B	=	Resource available for additional experiments
C_i	=	Cost of each additional experiment for variable X_i
D_i	=	Number of additional experiments for variable X_i

[1] Graduate Research Assistant, Department of Mechanical Engineering
[2] Post-Doctoral Researcher, Department of Mechanical Engineering
[3] Associate Professor, Department of Mechanical Engineering
[4] Materials Research Engineer, Materials and Manufacturing Directorate, Air Force Research Laboratory, Wright-Patterson AFB, OH 45433, USA

E_i	=	Number of initial data points for variable X_i
$gbest$	=	Global best of all pso particles from all iterations
k	=	Optimization iteration number
L	=	Lower bound at optimization variables
MCS	=	Monte Carlo Sampling
n_i	=	Number of total data points for variable X_i
N	=	Number of variables in the response function
$pbest_j$	=	Personal best of the jth pso particle until the current iteration
q_1	=	Individual weight
q_2	=	Social weight
r_{j1}	=	Random number between 0 and 1 for jth pso particle and first term
r_{j2}	=	Random number for jth pso particle and second term
S_i	=	Sample standard deviation for variable X_i
s_j^k	=	Position of jth pso particle at k iteration
t	=	Student's t-distribution
U	=	Upper bound at optimization variables
v_j^k	=	Velocity of jth pso particle at k iteration
v_j^{k+1}	=	Velocity of jth pso particle at k+1 iteration
w_j	=	Inertia weight of the jth pso particle
X_{E_i}	=	Values of X_i for E_i data points
X_{D_i}	=	Values of X_i for D_i additional experiments
X_i	=	Statistically independent variable in the response function
$\overline{X}_i$	=	Sample mean for variable X_i
Z	=	Response
χ^2	=	Chi-square distribution
μ_i	=	Population mean for variable X_i
μ_Z	=	Response mean
σ_i	=	Population standard deviation for variable X_i
σ_{μ_Z}	=	Standard deviation of the response mean
$\sigma_{\mu_Z} - orig$	=	Standard deviation of the response mean based on the original data
$\sigma_{\mu_Z} - opt$	=	Optimum solution of the standard deviation of the response mean
σ_Z	=	Response standard deviation

1. Introduction

It is well known that statistical estimates from simulation often involve significant uncertainty due to uncertainty in the input parameters, e.g. the probability-of-failure or the mean and standard deviation of the output. The input probability distributions may be developed from limited data, e.g., a few experiments, thus the mean and standard deviation of the inputs are also random variables dependent upon the amount of data available. The resulting uncertainty in the simulation outputs can be very large and should be accounted for in reporting the computational results and subsequent decision making.

Other authors have recognized this issue and presented strategies for determining the variation in the output results due to the uncertainty in the input PDF parameters. While the concept to accomplish this is straightforward by nesting the single output probability calculation (inner loop) inside another loop which accounts for the variation in the input parameters (outer loop), the practical accomplishment of such nested probability calculations is nontrivial. Several authors have developed computational strategies to address this issue. The strategy to date usually involved approximating the inner loop using a surrogate model (response surface), then calling the surrogate model from the outer loop. Thus, the analysis is efficient once the response surface has been constructed. However, the quality of the final solution will depend upon the quality of the surrogate model.

Cruse and Brown present a Bayesian analysis method to compute the confidence interval of the probability-of-failure using a directed acyclic graph approach [6]. The important random variables were first determined using sensitivity analysis (4 total), then these variables were used in a design of experiments study to construct the response surface. The conditional probability-of-failure as a function of the random variable input parameters was approximated using a polynomial response surface method using design of experiments.

Brown and Grandhi develop a response surface of the probability-of-failure with respect to the input PDF moments [2]. A derivative-enhanced Kriging model is used to develop the response surface. The computational complexity issue is then circumvented by replacing the inner loop with the response surface. The results show the Kriging model performs very well and confidence intervals of the probability-of-failure can be estimated in reasonable time. The results also show that the variation in the input parameters can cause dramatic variations in the probability-of-failure, resulting in very broad 95% confidence limits.

Ferson and Ginzburg emphasize that the nested approach (called second-order Monte Carlo in their paper) has inherent difficulties: (1) parameterization issues, (2) computational complexity, and (3) interpretational difficulty [5]. Parameterization requires the analyst to select how the parameters of the inputs vary which may introduce more uncertainty. However, in certain circumstances the distribution of the input parameters is well known theoretically. For example, the mean follows a t distribution and the standard deviation follows a chi-squared distribution [1]. Computational complexity is a continuing issue that requires clever strategies. The best approach to date is to develop a surrogate model of some form for the inner loop that defines the probabilistic output as a function of the parameters of the input PDFs. The interpretational approach recommended here is to compute confidence intervals (typically 95%).

While several authors have recognized the issue of uncertainty in the input parameter PDFs and the computational challenges that arise, less work has been done on assessing the resulting “action items” needed to reduce the confidence limits on the probability output(s). In this paper, an optimization method is combined with a simplified inner loop scheme to determine the subsequent experiments needed to most efficiently reduce the output variance subject to a cost (time, money, schedule) constraint. The action items considered in this work were to add additional data (experiments) to better characterize the mean and standard deviation of the input PDFs. In other words, what additional experiments should be conducted in order to improve the confidence in a probabilistic solution given cost constraints.

2. Methodology

A schematic of the computational approach is shown in Figure 1. The inner loop is shown simply as “Probabilistic Analysis”. The outer loop indicates that a Monte Carlo sampling method will be used to generate realizations of the mean and standard deviation of each random variable, evaluate the mean and standard deviation of a response, then repeat to develop statistics of the mean and standard deviation of the output response.

From statistical theory, the quantity $(\overline{X}-\mu)/(S/\sqrt{n})$ for an independent random variable has a Student’s t distribution with $n-1$ degrees of freedom, where $\overline{X}$ denotes the sample mean, μ the true population mean, S the sample standard deviation, and n the number of data points [1]. As a result, realizations of the population mean for random variable X_i that represent the expected variation in the population mean as a function of the number of data points are straightforward to determine as

$$\mu_i = \overline{X}_i + \frac{t_{n-1} S_i}{\sqrt{n_i}} \tag{1}$$

where t represents a realization from the standard t distribution.

Similarly, the quantity $(n-1)S^2/\sigma^2$ follows a chi-square distribution with $n-1$ degrees of freedom. A realization for the population standard deviation can be determined from

$$\sigma_i = \sqrt{\frac{(n-1)S_i^2}{\chi_{n-1}^2}} \tag{2}$$

where χ_{n-1}^2 is the chi-squared distribution with $n-1$ degrees of freedom.

A. Probabilistic Analysis

The determination of the probability-of-failure or the response moments is often computed using sampling (hence the term “inner loop”) or using other methods such as the First Order Reliability Method (FORM). Two primary cases where analytical methods can be use used are the normal format (linear combination of normal random variables) and the lognormal format (multiplicative combination of lognormal variables). The normal format is used in this paper to provide an efficient analysis method.

Given a response function $Z = A_0 + \sum_i^N A_i X_i$ where, A_0 and A_i are constants of the statistically independent normally distributed random variables X_i, the response mean and standard deviation can be computed exactly as

$$\mu_Z = A_0 + \sum A_i \mu_i \tag{3}$$

and

$$\sigma_Z = \sqrt{\sum A_i^2 \sigma_i^2} \tag{4}$$

B. Determination of Output Distribution

A schematic overview of the methodology is shown in Figure 1. The statistical process is explained as follows:

- A distribution of each variable X_i is provided.
- The population mean and standard deviation of each random variable is sampled from the t and chi-squared distribution, respectively, see Eqs. (1) and (2).
- The mean and standard deviation of the response are calculated as shown in Equations (3) and (4).
- Steps 2 and 3 are repeated many times to develop a population of response moments μ_Z and σ_Z
- Standard deviation of the response mean σ_{μ_Z} is calculated
- An optimization algorithm determines the optimal number of additional experiments D_i to minimize the standard deviation of the response moments subject to a constraint $\sum_i^N C_i D_i \leq B$, where C_i is the cost of each additional experiment and B is the total funds available

C. Optimization

The constraint optimization problem described above is formulated as following:

Objective: Minimize (σ_{μ_Z})

	Allocated resources, B
Constraint:	$\sum_{i}^{N} C_i D_i \le B$
Variable bounds:	$D_i^L < D_i < D_i^U$

There are many optimization methods available to solve an optimization problem. The presence of integer variables and high computational cost associated with the function evaluation suggests that a population-based approach is suitable for the problem. A particle swarm optimization (PSO) modified to handle integer variable is selected for the problem. PSO is selected over other population-based approach because of ease of implementation and lower number of user parameters.

PSO as proposed by Kennedy [3], modified the population from step-to-step based on a set of communication rules. In this set of rules, each particle uses its current fitness, its best fitness, the best fitness of all particles and a communication structure to determine movement parameters. These parameters change over the course of iterations, and particle population tends to converge and provide an optimum solution.

$$v_j^{k+1} = wv_j^k + q_1 r_{j1}(pbest_j - s_j^k) + q_2 r_{j2}(gbest - s_j^k) \quad (5)$$

Equation (5) calculates the velocity or movement of each particle for the k+1 iteration. This equation uses velocity (v_j^k), position (s_j^k), personal best ($pbest_j$), global best ($gbest$), and three constants (w, q_1, q_2) from iteration k to calculate the velocity of j^{th} particle. This velocity is used to calculate the new position of the particle as shown in Equation (6).

$$s_j^{k+1} = s_j^k + v_j^{k+1} \quad (6)$$

An investigation was performed to determine if a special strategy [7] was required to handle integer variables. A number of integer variable problems were solved to find out if rounding the velocity trajectory of the particles was sufficient to handle the integer variables. The results indicated that PSO was able to solve the problem and no special strategy was required to handle integer variables. Manual search was used to determine appropriate values of the optimization variable for each problem and once the parameters were decided, these parameters were fixed for all similar problems. For two and three variable numerical problems, the PSO parameters used are given in the following table.

Description (Symbols)	Value
Inertia weight (w_i)	0.529
Cognitive constant (q_1)	1.494

Social constant (q_2)	1.494
Maximum velocity	$\frac{D^U - D^L}{2}$
Minimum velocity	$-\frac{D^U - D^L}{2}$

The number of particles and the number of iterations were selected for each problem depending upon the number of variables and complexity of the problem. For all the problems in the paper, the optimization used normalized variables (between 0 and 1). This normalization assisted in optimizing problems with variables of different order of magnitudes and ranges.

The formulated problems were constrained optimization problems. The constraints were handled using a penalty function approach. The penalty function approach used here is a modified form of exterior penalty function method. In this approach, a feasible design point objective function value was always better than an infeasible design point. This was achieved by applying a huge penalty to the objective function value of an infeasible design point. This strategy forced all the particles to move toward the feasible design space.

The optimization algorithm employed two-convergence criteria to stop the optimization. The first criterion was that if the best solution does not change for 20 consecutive iterations then the optimization stops. The second criterion was that the optimization reaches the maximum number of iterations specified at the beginning of the optimization.

The problems solved in the paper involve a large number of random numbers (Monte Carlo samples, initial experiments, and additional experiments). This randomness is also involved while calculating the objective function values during optimization. The randomness is a challenging issue to obtain consist results from optimization because the randomness can change the optimum solution each time the same problem is solved. This is one of the main conclusions of this study. As described in the numerical examples, a range of studies were performed to investigate the effect of randomness in the optimization.

3. Numerical examples

The methodology described above was applied to two- and three-variable problems. All cases considered a linear response function and used the normal format to determine the response moments.

A. Two-variable problem

The response was defined by a linear function $Z = 1 + 10X_1 + 10X_2$, the variables X_1 and X_2 were normally distributed with μ_i and $\sigma_i$and E_i original data points were used. Each additional experiment had a cost of $1,000 and the total money available was $20,000. Therefore, the maximum additional number of experiments for each of the variables was 20.

A numerical study to determine a sufficient number of Monte Carlo samples was conducted using case i described below. The results indicated that convergence occurred at 10,000 samples. Therefore, 10,000 samples were used for all the case studies described below.

In the numerical examples below, the following notation is used.

- E_i represents the number of initial data points for random variable X_i,
- $X_{E\,i}$ represents the values of the initial data points for random variable X_i,
- D_i represents the number of additional experiments for random variable X_i,
- X_{D_i} represents the values of the additional experiments for random variable X_i,
- $\sigma_{\mu_Z} - orig$ represents the standard deviation of the mean of Z based on the original data $X_{E\,i}$
- $\sigma_{\mu_Z} - opt$ represents the optimum solution of the standard deviation of the mean of Z
- $X_{E\,i}$ and the additional experiments X_{D_i}

Several different cases were considered:

i $E_1 = E_2 = 10,\ X_{E_1} \neq X_{E_2},\ X_{D_i}\ \text{-}\ random$: That is, the number of initial experiments were set equal to 10 ($E_1 = E_2 = 10$), but the initial values were different ($X_{E_1} \neq X_{E_2}$). The results of the additional experiments (D_i) were allocated randomly according to the input distribution, e.g., normal with mean and standard deviation equal to 10 and 3, respectively.

ii $E_1 = E_2 = 10,\ X_{E_1} \neq X_{E_2},\ X_{D_1} = X_{D_2} = \mu$: Similar to case i except the results for any additional experiments were set at the mean values.

iii $E_1 = E_2 = 10,\ X_{E_1} = X_{E_2},\ X_{D_i}\ \text{-}\ random$: Similar to case i except the initial experimental results were the same and allocated randomly according to the input distributions

iv $E_1 = E_2 = 10,\ X_{E_1} = X_{E_2},\ X_{D_1} = X_{D_2} = \mu$: The values for the initial experiments are the same and the results of any new experiments set at the mean of the distributions.

v $E_1 = 20,\ E_2 = 10,\ X_{E_1} \neq X_{E_2},\ X_{D_i}\ \text{-}\ random$: A different number of initial data points and values were used for X_1 and X_2. The additional experiments were allocated randomly according to the input distribution.

vi $E_1 = 20,\ E_2 = 10,\ X_{E_1} \neq X_{E_2},\ X_{D_1} = X_{D_2} = \mu$: Similar to case v except the values for the additional experiments were located at mean of the variable.

Case i: $E_1 = E_2 = 10,\ X_{E_1} \neq X_{E_2},\ X_{D_i}\ \text{-}\ random$

The parameters are defined in Table 1 and the design space is shown in Figure 2; the colored area indicates the feasible area while the flat blue area represents the infeasible area. The values for the initial data points, X_{E_i}, are shown in Table 2. The values for the results of the additional experiments, X_{D_i}, were determined randomly according to the random variable distribution. For example, the values chosen for X_{D_1} were obtained randomly from a normal distribution with mean and standard deviation of 10 and 3, respectively. Because of this randomness, different results were obtained for D_i for each analysis.

The optimization process was conducted using 10 iterations and 20 particles. Each addition experiment cost \$1000 and the total amount funds available was \$20,000. The particles during the 1st, 5th and 10th iteration for one particular analysis is shown in Figure 3. Table 3 show the results obtained from running the analysis 5 separate times. The effect of the randomness in the values of X_E and X_D is evident; a large variation in the optimal number of additional experiments for each variable was obtained. The problem appears symmetric with respect to X_1 and X_2; therefore, the intuitive solution is $D_1 = D_2 = 10$. However, due to the random nature of the actual values used during the analysis, quite different answers can be obtained.

In all cases, the additional experiments reduced the standard deviation of μ_z by approximately 30%. For example, in run 1, the standard deviation of the response mean without any additional data was 18.28 and was reduced to 11.23 after adding the additional experiments.

The PDF of the response mean is shown in Figure 4 for run 1. The red dotted line shows the PDF of the response mean previous to any additional data, the solid blue line shows the PDF after adding additional experiments. A reduction in the standard deviation of μ_z is evident. Note, also that due to the random allocation of the results from the additional experiments, the mean after adding the additional experiments can shift from the original data.

Table 1. Variable Parameters (Case i)

	A_i	μ_{X_i}	σ_{X_i}	E_i	C_i
X_1	10	10	3	10	\$1,000
X_2	10	10	3	10	\$1,000

Table 2. Values for variables X_{E_i} and X_{D_i} (Case i)

X_{E_1}	X_{E_2}
9.81	9.25
14.33	9.41
5.89	11.10
5.38	14.44
7.09	7.26
8.60	10.04
10.78	8.73
11.80	9.34
16.06	11.31
12.62	11.47

X_{D_1}	X_{D_2}
9.14	11.71
9.18	11.57
9.04	10.76
11.19	8.83
8.61	11.86
10.52	7.85
2.05	8.98
8.46	6.98
5.39	6.94
6.88	5.81
9.31	9.05
5.13	8.95

5.91	12.35
7.62	12.30
10.86	11.16
11.32	11.41
7.86	8.44
8.20	11.43
11.55	11.04
5.84	11.43

Table 3. Summary (Case i)

Run	D_1	D_2	$\sigma_{\mu_z} - orig$	$\sigma_{\mu_z} - opt$	% Reduction	$\Sigma C_i D_i$
1	15	5	18.28	11.23	38.6%	\$20,000
2	16	4	20.46	13.68	33.1%	\$20,000
3	6	14	19.18	12.69	33.8%	\$20,000
4	11	9	21.61	14.83	31.4%	\$20,000
5	7	13	25.17	16.76	33.4%	\$20,000

Case ii: $E_1 = E_2 = 10,\ X_{E_1} \neq X_{E_2},\ X_{D_1} = X_{D_2} = \mu$

For this case, the additional experiments, X_{D_i} were set the mean of each variable, that is, it was assumed that after conducting the new experiments, the results were equal to the population mean. Clearly, this is an assumption not realized in practice but useful to examine the affect of randomness in the optimization process.

The design space is shown in Figure 5, the colored area shows the feasible area, while the flat blue area represents the infeasible area. The surface shown in Figure 5 is smoother than in case i due to the removal of some randomness. The optimization process used 10 iterations with 10 particles. PSO require lower number of particles as compared to Case i because the removal of randomness provides a monotonous (smooth) design space.

The results for 5 analyses are shown in Table 6. The results are a little more symmetric than case i, as expected, but still show a significant amount of asymmetry. The optimization converged after a combination of 10 iterations and 10 particles and the optimization process is represented in Figure 6. The addition of the new experimental data reduced the standard deviation of the response mean by approximately 50% in each run. The PDF of the response mean is shown in Figure 7. The red dotted line shows the PDF of the response mean previous to any additional data, the solid blue line shows the PDF after adding a new set of data.

Table 4. Variable Parameters (Case ii)

	A_i	μ_{X_i}	σ_{X_i}	E_i	C_i
X_1	10	10	3	10	\$1,000
X_2	10	10	3	10	\$1,000

Table 5. Values for variables X_{E_i} and X_{D_i} (Case ii)

X_{E_1}	X_{E_2}
9.81	9.25
14.33	9.41
5.89	11.10
5.38	14.44
7.09	7.26
8.60	10.04
10.78	8.73
11.80	9.34
16.06	11.31
12.62	11.47

X_{D_1}	X_{D_2}
10	10
10	10
10	10
10	10
10	10
10	10
10	10
10	10
10	10
10	10
10	10
10	10
10	10
10	10
10	10
10	10
10	10
10	10
10	10
10	10

Table 6. Summary (Case ii)

Run	D_1	D_2	$\sigma_{\mu_Z} - orig$	$\sigma_{\mu_Z} - opt$	% Reduction	$\Sigma C_i D_i$
1	13	7	18.28	8.51	53.4%	\$20,000
2	13	7	20.46	10.43	49.0%	\$20,000
3	9	11	19.18	9.37	51.1%	\$20,000
4	10	10	21.61	11.33	47.6%	\$20,000
5	11	9	25.17	12.61	49.9%	\$20,000

Case iii: $E_1 = E_2 = 10,\ X_{E_1} = X_{E_2},\ X_{D_i}$ *- random*

For this case, the initial values of the variables are equal, $X_{E_1} = X_{E_2}$, and the values of the additional experiments were determined randomly according to the random variable distributions. The problem parameters (A_i, μ_{X_i}, σ_{X_i}, E_i, C_i) are shown in Table 7 and the experimental location values (X_{E_i}, X_{D_i}) are shown in Table 8. The design space is shown in Figure 8 where the colored area shows the feasible area, while the flat blue area represents the infeasible area. 10 iterations and 20 particles were used during the optimization process. The particle movement for first, fifth and tenth iteration is shown in Figure 9.

The asymmetry in the allocation of the additional experiments seemed to be reduced compared to both cases i and ii; however, more analyses are needed to confirm this conclusion. The addition of the new experimental data reduced the standard deviation of the response mean by close to 40%.

The PDF of the response mean is shown in Figure 10. The red dotted line shows the PDF of the response mean previous to any additional data, the solid blue line shows the PDF after adding a new set of data.

Table 7. Variable Parameters (Case iii)

	A_i	μ_{X_i}	σ_{X_i}	E_i	C_i
X_1	10	10	3	10	$1,000
X_2	10	10	3	10	$1,000

Table 8. Values for variables X_{E_i} and X_{D_i} (Case iii)

X_{E_1}	X_{E_2}
9.96	9.96
8.26	8.26
16.41	16.41
9.23	9.23
5.77	5.77
15.31	15.31
10.98	10.98
6.64	6.64
11.86	11.86
13.81	13.81

X_{D_1}	X_{D_2}
12.51	8.04
7.83	6.76
7.84	9.86
9.40	11.14
9.94	9.01
10.84	8.50
13.17	9.89
11.87	9.48
4.75	7.13
12.09	13.88
12.43	11.32
11.91	13.84
13.93	8.51
10.98	6.64
7.98	12.42
9.55	10.12
2.65	7.73
11.42	9.73
10.35	3.97
8.23	13.25

Table 9. Summary (Case iii)

Run	D_1	D_2	$\sigma_{\mu_Z} - orig$	$\sigma_{\mu_Z} - opt$	% Reduction	$\Sigma C_i D_i$
1	8	12	23.34	13.53	42.0%	$20,000
2	14	6	19.61	13.53	31.0%	$20,000
3	10	10	22.88	14.26	37.7%	$20,000
4	9	11	21.47	13.12	38.9%	$20,000
5	10	10	22.88	14.26	37.7%	$20,000

Case iv: $E_1 = E_2 = 10,\ X_{E_1} = X_{E_2},\ X_{D_1} = X_{D_1} = \mu$

For this case, the initial values of the variables were equal, $X_{E_1} = X_{E_2}$, and the values of the additional experiments were also equal and set at the mean value, $X_{D_1} = X_{D_2} = \mu$ determined according to the random variable distribution. The parameters are shown in Table 10 and Table 11. 10 iterations and 10 particles were used. The design space is shown in Figure 11 where the colored area shows the feasible area, while the flat blue area represents the infeasible area.

The additional experiments of each variable were 10 and 10 respectively, as expected; in this case, there is no asymmetry. The optimization converged after a combination of 10 iterations and 10 particles and the optimization process is represented in Figure 12. The addition of the new experimental data reduced the standard deviation of the response mean from a value of 23.34 to a value of 11.53; a reduction of 50.6%.

The PDF of the response mean is shown in Figure 13. The red dotted line shows the PDF of the response mean previous to any additional data, the solid blue line shows the PDF after adding a new set of data. The standard deviation of the response mean was reduced by approximately 50%.

Table 10. Variable Parameters (Case iv)

	A_i	μ_{X_i}	σ_{X_i}	E_i	C_i
X_1	10	10	3	10	$1,000
X_2	10	10	3	10	$1,000

Table 11. Values for variables X_{E_i} and X_{D_i} (Case iv)

X_{E_1}	X_{E_2}
9.96	9.96
8.26	8.26
16.41	16.41
9.23	9.23
5.77	5.77
15.31	15.31
10.98	10.98
6.64	6.64
11.86	11.86
13.81	13.81

X_{D_1}	X_{D_2}
10	10
10	10
10	10
10	10
10	10
10	10
10	10
10	10
10	10
10	10
10	10
10	10
10	10
10	10
10	10
10	10
10	10
10	10
10	10
10	10

Table 12. Summary (Case iv)

Run	D_1	D_2	$\sigma_{\mu_z} - orig$	$\sigma_{\mu_z} - opt$	% Reduction	$\Sigma C_i D_i$
1	10	10	23.34	11.53	50.6%	$20,000
2	10	10	14.86	7.3	50.9%	$20,000
3	10	10	15.38	7.87	48.8%	$20,000
4	10	10	12.36	6.05	51.1%	$20,000
5	10	10	16.96	8.41	50.4%	$20,000

Case v: $E_1 = 20,\ E_2 = 10,\ X_{E_1} \neq X_{E_2},\ X_{D_i}$ *- random*

For this case, the number of initial experiments and the values of the experiments were not equal; and the numbers of initial data points were 20 and 10 respectively, and the values of the additional experiments were determined according to the random variable distributions. The parameters are shown in Table 13 and the design space is show in Figure 14 where the colored area shows the feasible area, while the flat blue area represents the infeasible area. The optimization used 10 iterations and 20 particles. The optimization process is represented in Figure 15 for run 1.

In this case, the number of additional experiments for each variable was heavily weighted towards more experiments of X_2, see Table 15. This occurs largely because more information is already available for X_1; therefore, the variance of the population mean μ_1 is tighter than μ_2. The addition of the new experimental data reduced the standard deviation of the response mean by 20-40%.

The PDF of the response mean is shown in Figure 16. The red dotted line shows the PDF of the response mean previous to any additional data, the solid blue line shows the PDF after adding a new set of data.

Table 13. Variable Parameters (Case v)

	A_i	μ_{X_i}	σ_{X_i}	E_i	C_i
X_1	10	10	3	20	$1,000
X_2	10	10	3	10	$1,000

Table 14. Values for variables X_{E_i} and X_{D_i} (Case v)

X_{E_1}	X_{E_2}
11.60	7.91
14.10	18.44
7.73	12.86
12.34	7.45
9.31	10.81
12.46	12.02
7.68	12.16
8.00	10.75
13.68	9.31
8.54	13.77
4.72	
12.68	
9.32	
6.81	
10.23	
9.01	
4.53	
5.86	
12.73	
7.61	

X_{D_1}	X_{D_2}
11.80	8.03
5.29	8.32
13.69	12.63
9.20	8.00
6.38	10.83
7.17	4.11
13.95	7.40
5.03	9.37
9.63	5.62
9.03	5.13
16.91	9.96
9.32	14.06
12.42	9.70
9.30	15.27
9.25	12.82
12.60	9.31
13.45	10.98
4.78	7.25
9.70	11.78
8.78	10.11

Table 15. Summary (Case v)

Run	D_1	D_2	$\sigma_{\mu_Z} - orig$	$\sigma_{\mu_Z} - opt$	% Reduction	$\Sigma C_i D_i$
1	1	19	16.99	12.49	26.5%	$20,000
2	3	17	13.96	11.18	19.9%	$20,000
3	1	18	15.77	10.38	34.2%	$19,000
4	6	13	17.94	12.05	32.8%	$19,000
5	1	19	19.98	11.54	42.2%	$20,000

Case vi: $E_1 = 20,\ E_2 = 10,\ X_{E_1} \neq X_{E_2},\ X_{D_1} = X_{D_1} = \mu$

For this case, the number of initial experiments and the values of the experiments were not equal, the numbers of initial data points were 20 and 10 respectively, and the values of the additional experiments were set at the mean value of the random variables. The parameters are shown in Table 16 and the design space is show in Figure 17 where the colored area shows the feasible area, while the flat blue area represents the infeasible area. For this case, the number of additional experiments were weighted towards X_2 but not as heavily as in case v, see Table 18. The optimization converged after a combination of 10 iterations and 10 particles and the optimization process is represented in Figure 18. The addition of the new experimental data reduced the standard deviation of the response mean by approximately 40%.

The PDF of the response mean is shown in Figure 19. The red dotted line shows the PDF of the response mean previous to any additional data, the solid blue line shows the PDF after adding a new set of data.

Table 16. Variable Parameters (Case vi)

	A_i	μ_{X_i}	σ_{X_i}	E_i	C_i
X_1	10	10	3	20	$1,000
X_2	10	10	3	10	$1,000

Table 17. Values for variables X_{E_i} and X_{D_i} (Case vi)

X_{E_1}	X_{E_2}
11.60	7.91
14.10	18.44
7.73	12.86
12.34	7.45
9.31	10.81
12.46	12.02
7.68	12.16
8.00	10.75
13.68	9.31
8.54	13.77
4.72	
12.68	
9.32	
6.81	
10.23	
9.01	

X_{D_1}	X_{D_2}
10	10
10	10
10	10
10	10
10	10
10	10
10	10
10	10
10	10
10	10
10	10
10	10
10	10
10	10
10	10
10	10

4.53			10	10
5.86			10	10
12.73			10	10
7.61			10	10

Table 18. Summary (Case vi)

Run	D_1	D_2	$\sigma_{\mu_Z} - orig$	$\sigma_{\mu_Z} - opt$	% Reduction	$\Sigma C_i D_i$
1	6	14	16.99	9.6	43.5%	\$20,000
2	13	7	13.34	8.76	34.3%	\$20,000
3	6	14	19.28	10.43	45.9%	\$20,000
4	8	12	21.1	11.93	43.5%	\$20,000
5	7	13	18.49	10.16	45.1%	\$20,000

B. Three random variables

The response was define by a linear function $Z = 1 + 7X_1 + 3X_2 + 13X_3$, the variables X_1, X_2 and X_3 were normally distributed. Different values were used for A's, input PDF parameters, initial data points and cost functions. The total funds available was \$20,000. Two different cases were considered:

i The results of the additional experiments were allocated randomly according to the input distribution.

ii The results of the additional experiments are allocated at the mean of the variable.

Case vii: $E_1 = 5,\ E_2 = 15,\ E_3 = 10,\ X_{E_1} \neq X_{E_2} \neq X_{E_3},\ X_{D_i}$ *- random*

The parameters are described in Table 19 and Table 20, the values for the results of the additional experiments were determined according to the random variable distribution.

Table 19. Variable Parameters (Case vii-Three variables)

	A_i	μ_{X_i}	σ_{X_i}	E_i	C_i
X_1	7	10	3	5	\$700
X_2	3	0	1	15	\$500
X_3	13	7	2	10	\$1,000

Table 20. Values for variables X_{E_i} and X_{D_i} (Case vii-Three variables)

X_{E_1}	X_{E_2}	X_{E_3}
9.07	-0.35	9.07
11.64	-0.58	6.57
13.93	0.14	5.12
14.92	-0.79	7.10
14.69	-0.14	10.62
	-0.15	7.15

X_{D_1}	X_{D_2}	X_{D_3}
9.30	1.51	9.30
7.33	-0.52	9.68
11.75	1.75	10.32
12.61	1.10	9.79
14.71	-0.02	8.51
8.47	-0.28	6.67

-0.21	7.18
-0.41	5.68
0.15	9.15
0.47	7.51
1.76	
-0.41	
1.00	
0.56	
-0.57	

12.40	1.33	6.18
11.85	0.46	4.16
12.21	-1.42	6.84
8.89	0.39	6.57
12.69	0.15	7.51
9.53	1.06	7.46
9.05	1.03	7.52
8.77	-0.40	6.19
11.43	-1.16	6.64
15.48	-1.80	6.28
8.97	-0.17	5.99
9.44	0.88	8.98
4.78	0.29	9.98
11.26	0.76	5.16
11.93	-2.03	
12.80	0.04	
8.18	-0.13	
8.97	0.40	
7.45	1.47	
13.13	-0.10	
11.29	-0.73	
10.40	1.52	
	-0.45	
	-2.05	
	-1.04	
	-1.74	
	0.62	
	0.47	
	-0.24	
	0.31	
	-0.52	
	-0.34	
	-0.26	
	1.43	

The following study was performed to investigate the behavior of the problem. The numbers of additional experiments were determined using a simple allocation scheme. For example, equal money for each variable or maximum allocation for each variable, the summary is shown in Table 21. Thus, Table 21 show the results that can be expected without optimization.

Table 21. Summary of study (Case vii-Three variables)

D_1	28	0	0	9	9
D_2	0	40	0	13	9
D_3	0	0	20	7	9
$\sigma_{\mu_z} - opt$	10.87	15.34	12.56	10.59	10.77
% Reduction	30%	1%	19%	31%	30%
$\Sigma C_i D_i$	19600	20000	20000	19800	19800

The results from the optimization process are shown in Table 22. The standard deviation of the response mean was reduced by typically 40%. The PDF of the response mean is shown in Figure 20 for run 1. 20 iterations and 40 particles were used. The red dotted line shows the PDF of the response mean previous to any additional data, the solid blue line shows the PDF after adding a new set of data. Due to the random allocation of the results from the additional experiments, the mean after adding the additional experiments shifted to the right.

Table 22. Summary (Case vii-Three variables)

Run	D_1	D_2	D_3	$\sigma_{\mu_Z} - orig$	$\sigma_{\mu_Z} - opt$	% Reduction	$\Sigma C_i D_i$
1	18	0	7	15.45	9.61	38%	\$19,600
2	10	0	13	19.25	10.06	48%	\$20,000
3	11	0	12	26.22	14.03	46%	\$19,700
4	17	0	8	21.62	11.12	49%	\$19,900
5	9	1	13	13.67	11.39	17%	\$19,800

Whit the same parameters set in Table 22, the standard deviation of the response mean is calculated by changing the additional number of experiments by a small number and compared to the optimum solution of 9.61. This study was performed to determined the robustness of the solution. Table 23 shows different possibilities.

Table 23. Study around the optimal solution (Case vii-Three variables)

D_1	17	18	17	16	15
D_2	0	0	0	0	0
D_3	7	6	8	8	9
$\sigma_{\mu_Z} - opt$	9.72	9.74	9.86	10.22	9.84
% Increment	1%	1%	3%	6%	2%
$\Sigma C_i D_i$	18900	18600	19900	19200	19500

Case viii: $E_1 = 5,\ E_2 = 15,\ E_3 = 10,\ X_{E_1} \neq X_{E_2} \neq X_{E_3};\ X_{D_1} = 10,\ X_{D_2} = 0,\ X_{D_3} = 7$

The parameters are described in Table 24, the values for the results of the additional experiments were set at the mean. The optimization process obtained an optimal number of additional experiments of as shown in Table 26. 20 iterations and 40 particles were used. The standard deviation of the response mean was reduced by approximately 60%. The PDF of the response mean is shown in Figure 21. The red dotted line shows the PDF of the response mean previous to any additional data, the solid blue line shows the PDF after adding a new set of data

Table 24. Variable Parameters (Case viii-Three variables)

	A_i	μ_{X_i}	σ_{X_i}	E_i	C_i
X_1	7	10	3	5	\$700
X_2	3	0	1	15	\$500
X_3	13	7	2	10	\$1,000

Table 25. Values for variables X_{E_i} and X_{D_i} (Case viii-Three variables)

X_{E_1}	X_{E_2}	X_{E_3}	X_{D_1}	X_{D_2}	X_{D_3}
9.07	-0.35	9.07	10	0	7
11.64	-0.58	6.57	10	0	7
13.93	0.14	5.12	10	0	7
14.92	-0.79	7.10	10	0	7
14.69	-0.14	10.62	10	0	7
	-0.15	7.15	10	0	7
	-0.21	7.18	10	0	7
	-0.41	5.68	10	0	7
	0.15	9.15	10	0	7
	0.47	7.51	10	0	7
	1.76		10	0	7
	-0.41		10	0	7
	1.00		10	0	7
	0.56		10	0	7
	-0.57		10	0	7
			10	0	7
			10	0	7
			10	0	7
			10	0	7
			10	0	7
			10	0	
			10	0	
			10	0	
			10	0	
			10	0	
			10	0	
			10	0	
			10	0	
				0	
				0	
				0	
				0	
				0	
				0	
				0	
				0	
				0	
				0	
				0	
				0	

Table 26. Summary (Case viii-Three variables)

Run	D_1	D_2	D_3	$\sigma_{\mu_z} - orig$	$\sigma_{\mu_z} - opt$	% Reduction	$\Sigma C_i D_i$
1	17	0	8	15.47	6.89	56%	$19,900
2	14	0	10	14.95	5.91	60%	$19,800
3	14	0	10	18.24	7.31	60%	$19,800
4	20	0	6	22.65	7.4	67%	$20,000
5	14	0	10	22.19	8.49	62%	$19,800

6	14	0	10	16.53	6.3	62%	$19,800

4. Conclusions

Several simple probabilistic models were explored to assess the optimum additional experiments to reduce the output variance in the mean or standard deviation of a probabilistic response. The example problems were linear combinations of normal random variables such that the inner loop could be computed in closed form to facilitate efficient exploration; however, methodology is not limited to such simple problems. An efficient optimization strategy was implemented using a particle swarm optimization (PSO) method modified to handle integer variables. PSO was selected over other population-based approaches because of ease of implementation and a lower number of user parameters.

The number of additional experiments to add for each random variable to reduce the standard deviation of the response mean depend upon many factors: the number of initial data points (E_i), the importance of the random variable in the response (A_i), the range of the mean of the random variable (μ_i), the cost of each experiment (C_i). In addition, the actual values affect the solution. For complicated problems, it is not possible to ascertain an optimal allocation of additional experiments, hence the value of the approach described here.

The results indicate that the optimization is difficult due primarily to the randomness in the results of the additional experiments. The optimization results were dependent upon how the additional experiments were chosen. If chosen randomly, the optimum experiments were also random, that is, different results were obtained if the same analysis was rerun; however, general trends were observed. More consistent results were obtained if the additional experiments were all located at the mean value of the input random variables. This was evident in the symmetric two random variable problem where the intuitive results that and equal number of both experiments were needed would often not be obtained unless the initial and additional experimental results were selected to be the same for both variables. An examination of the three random variable problem clearly showed the value of the optimization algorithm since a larger reduction in the standard deviation of the response mean was obtained compared to ad hoc approaches.

5. References

[1] Ang, A.H.-S. and H.T. Wilson, *Probability Concepts in Engineering: Emphasis on Applications to Civil and Environmental Engineering*. 2nd Edition ed. 2007: John Wiley & Sons, Inc. 406.

[2] Brown, J.M. and R.V. Grandhi, *Probabilistic Gradient Kriging to Efficiently Predict Failure Probability Confidence Intervals*, in *49th AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference*. 2008: Schaumburg, IL. AIAA 2008-1715

[3] Kennedy, J. and R. C. Eberhart, *Particle Swarm Optimization*. Proceeding of IEEE conference on Neural Network, IV, Piscataway, NJ, 1995. p 1942-1948

[4] Walpole, R.E. and R.H. Myers, *Probability and Statistics for Engineers and Scientists*. Fifth Edition ed. 1993, New York: Macmillan Publishing Company.

[5] Ferson, S. and L.R. Ginzburg, Different methods are needed to propagate ignorance and variability, Reliability Engineering and System Safety 54 (1996) 133-144

[6] Cruse, T.A. and J.M. Brown, *Confidence Interval Simulation for Systems of Random Variables.* Journal of Engineering for Gas Turbines and Power, 2007. **129**(3): p. 836-842.

[7] G. Singh and R. Grandhi, *Mixed-Variable Optimization Strategy Employing Multifidelity Simulatuion and Surrogate Models*. AIAA Journal, 2010. **48**(1): p.215-223

[8] Venkataraman, S., et al., *Calculating Confidence Bounds for Reliability Index to Quantify Effect of Distribution Parameter Uncertainty*, in *48th AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference*. 2007: Honolulu, Hawaii. p. 19. AIAA 2007-1940

6. Figures

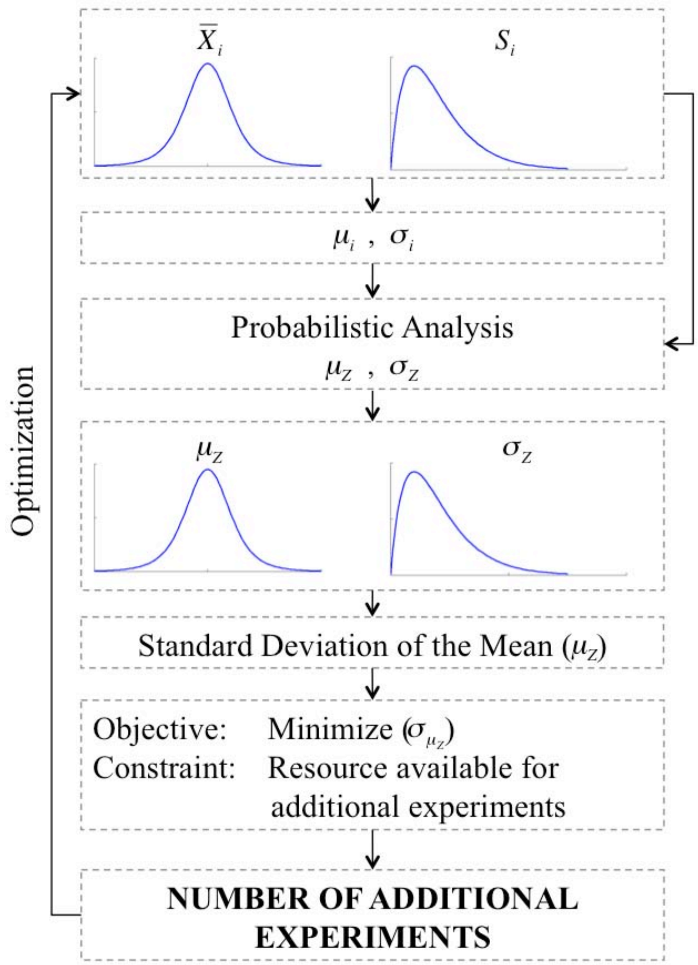

Figure 1. Schematic flow chart of the optimal allocation methodology

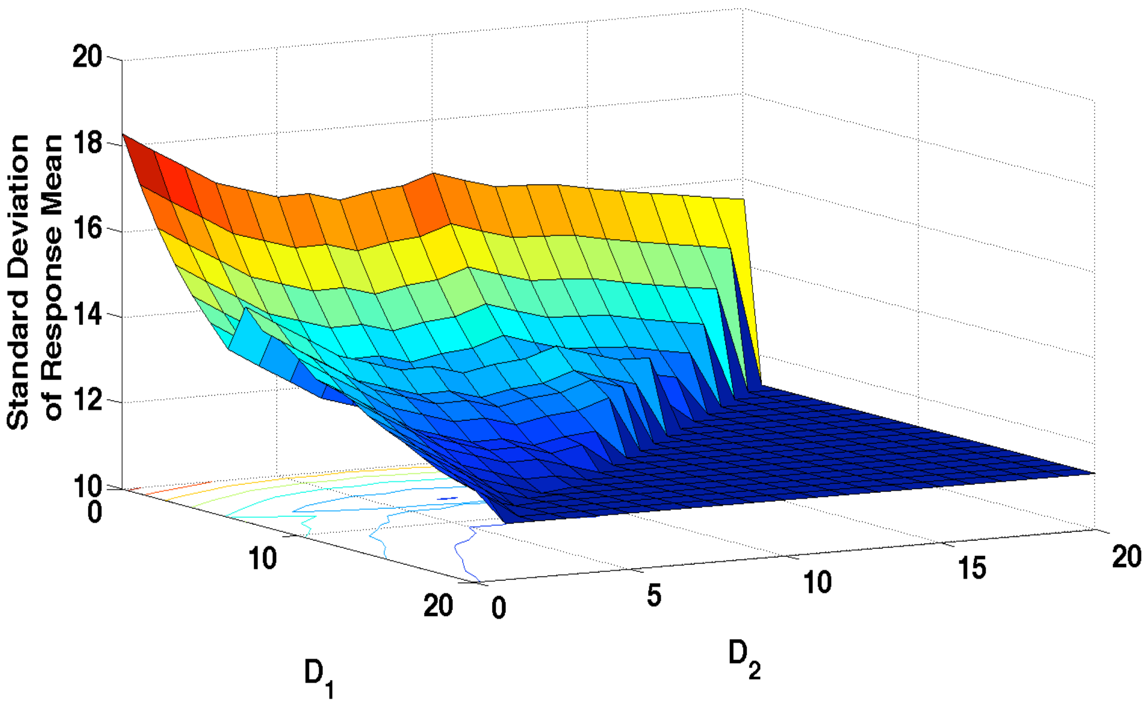

Figure 2. Design Space (Two variable - Case i)

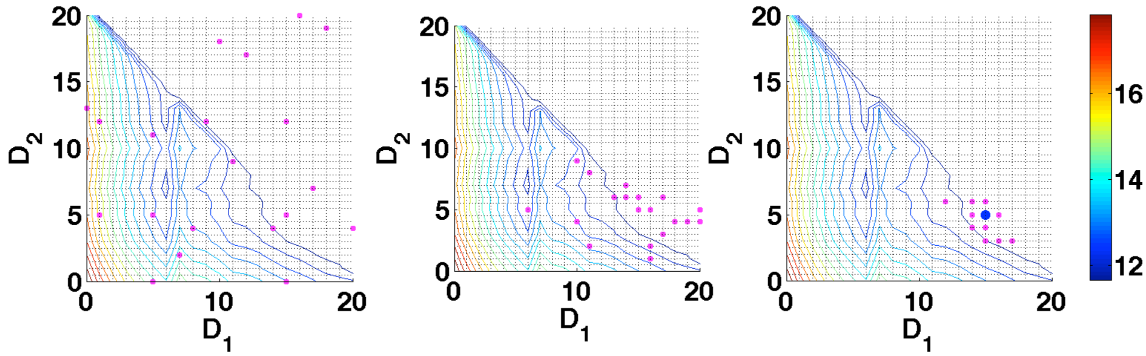

Figure 3. Particle movement (Two variable - Case i) a) Particles at first iteration, b) Particle at fifth iteration and c) Particles at the last iteration where the blue dot represents the optimum solution.

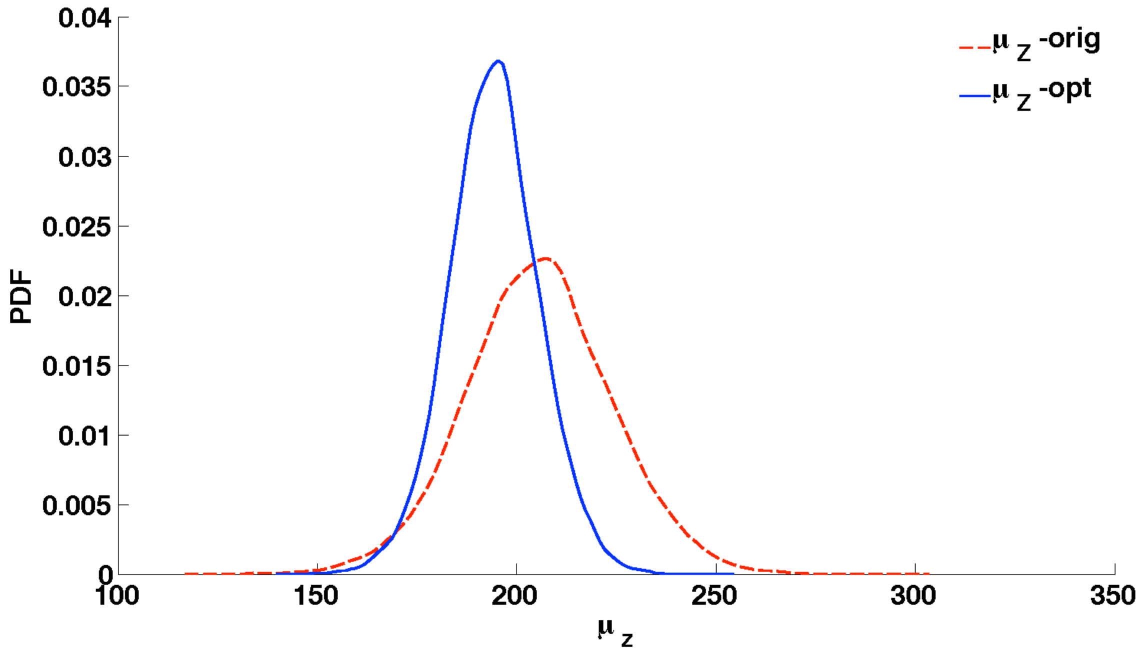

Figure 4. **PDF of Response mean (Two variables – Case i)**

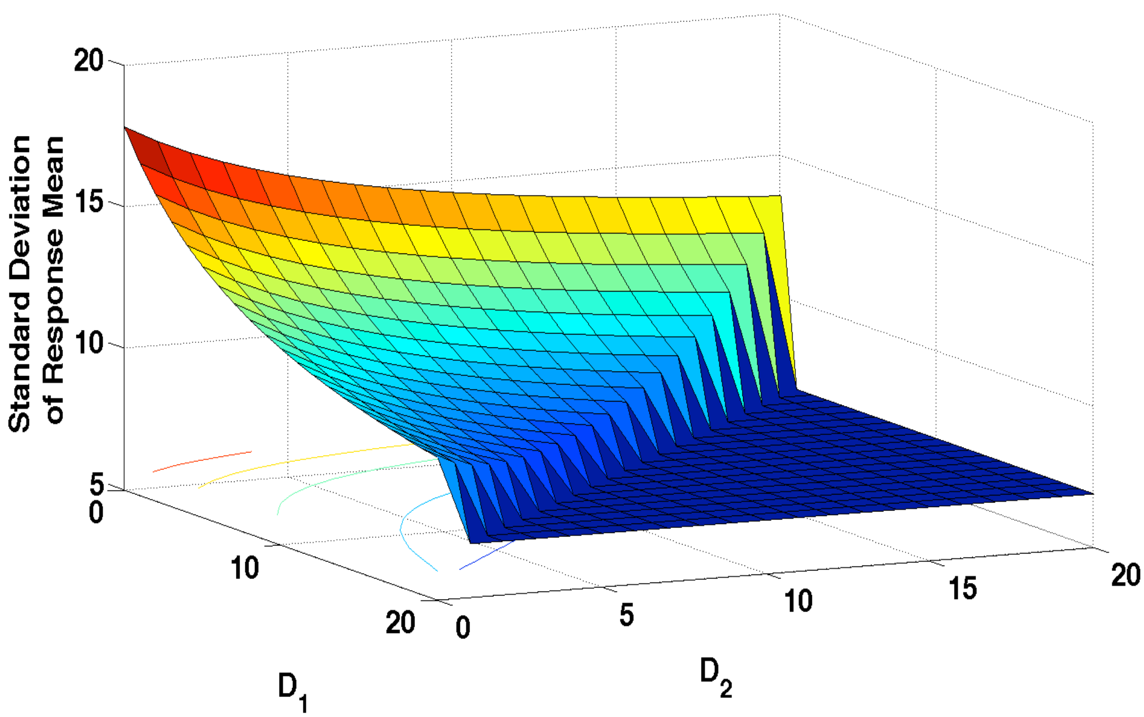

Figure 5. **Design Space (Two variable – Case ii)**

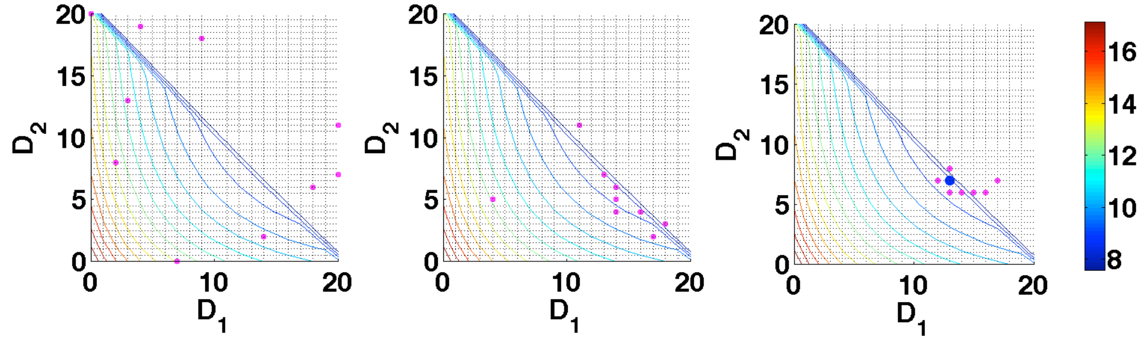

Figure 6. Particle movement (Case ii) a) Particles at first iteration, b) Particle at fifth iteration and c) Particles at the last iteration where the blue dot represents the optimum solution.

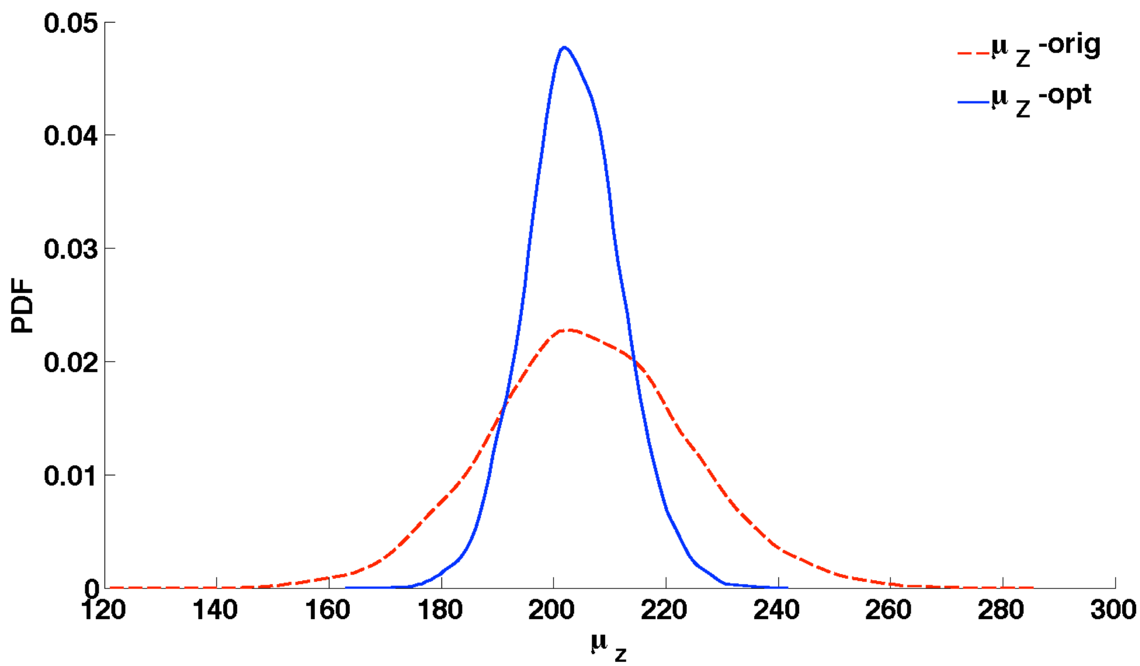

Figure 7. PDF of Response mean (Two variables – Case ii)

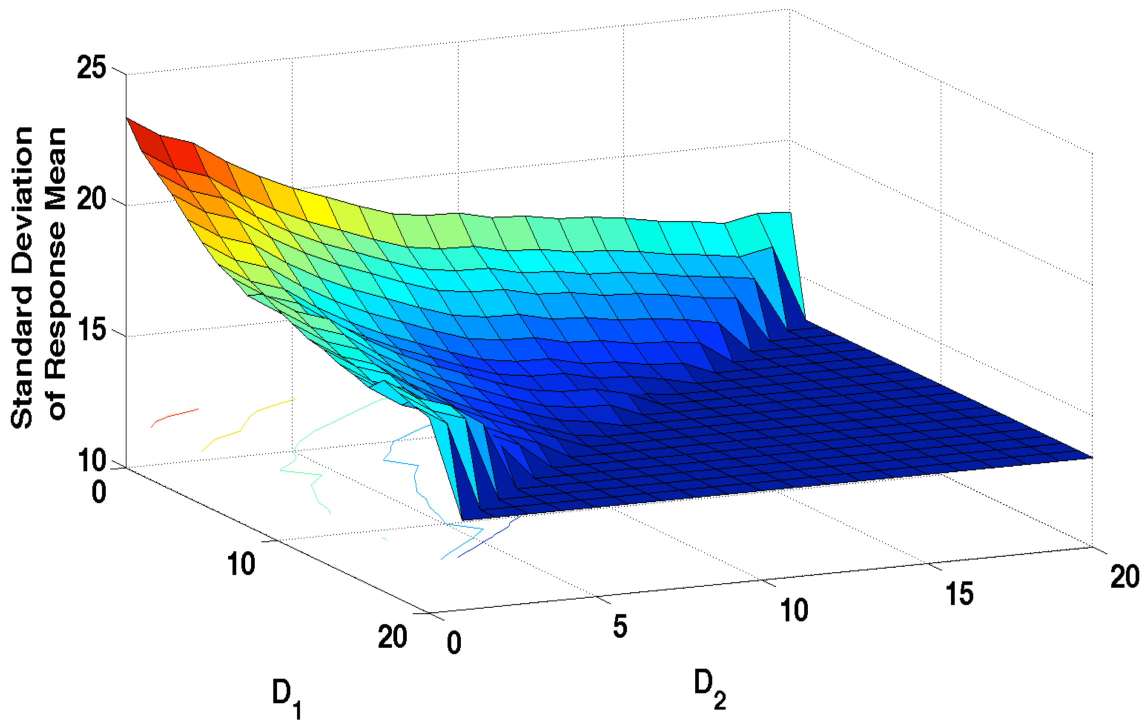

Figure 8. **Design Space (Two variable - Case iii)**

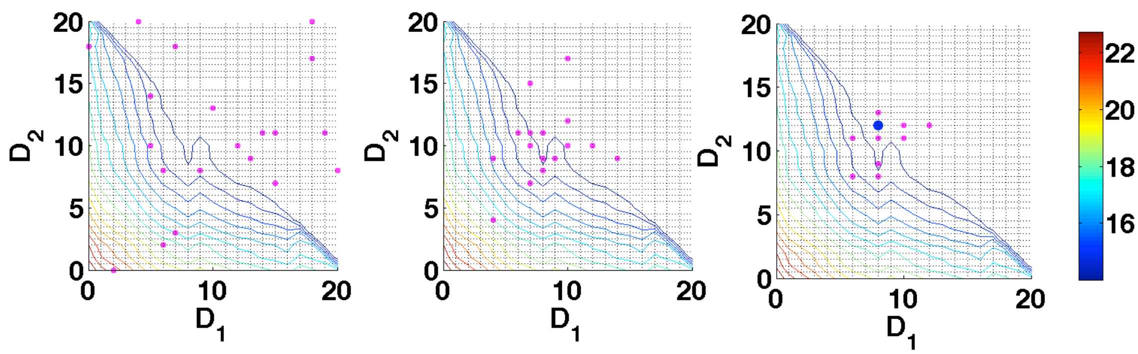

Figure 9. **Particle movement (Case iii)** a) Particles at first iteration, b) Particle at fifth iteration and c) Particles at the last iteration where the blue dot represents the optimum solution.

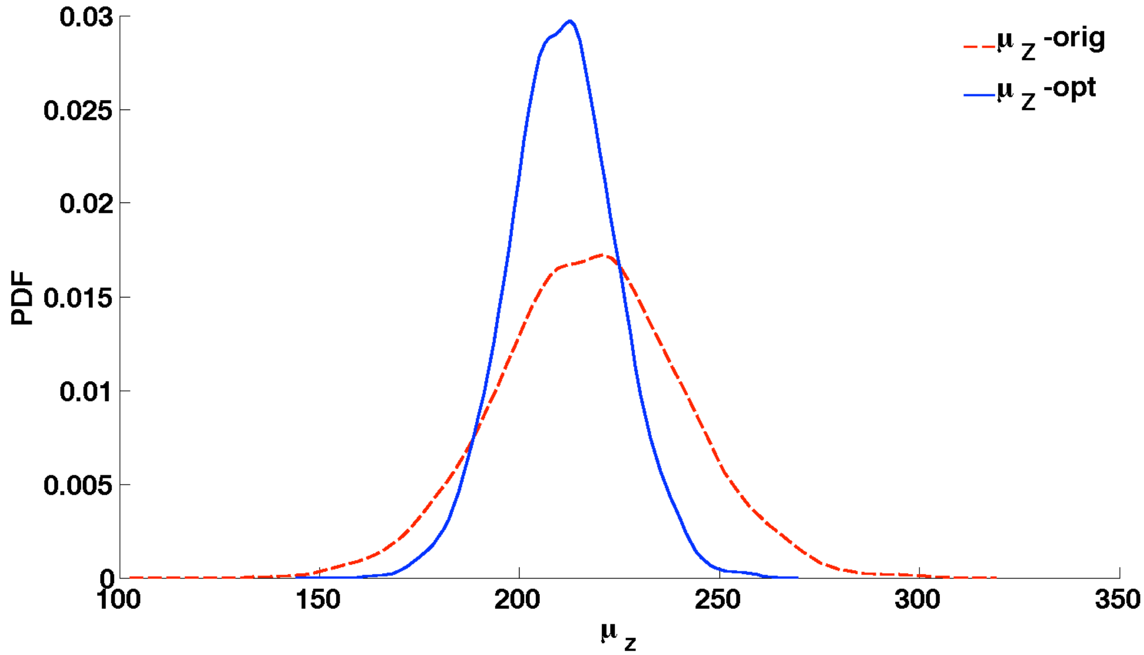

Figure 10. PDF of Response mean (Two variables – Case iii)

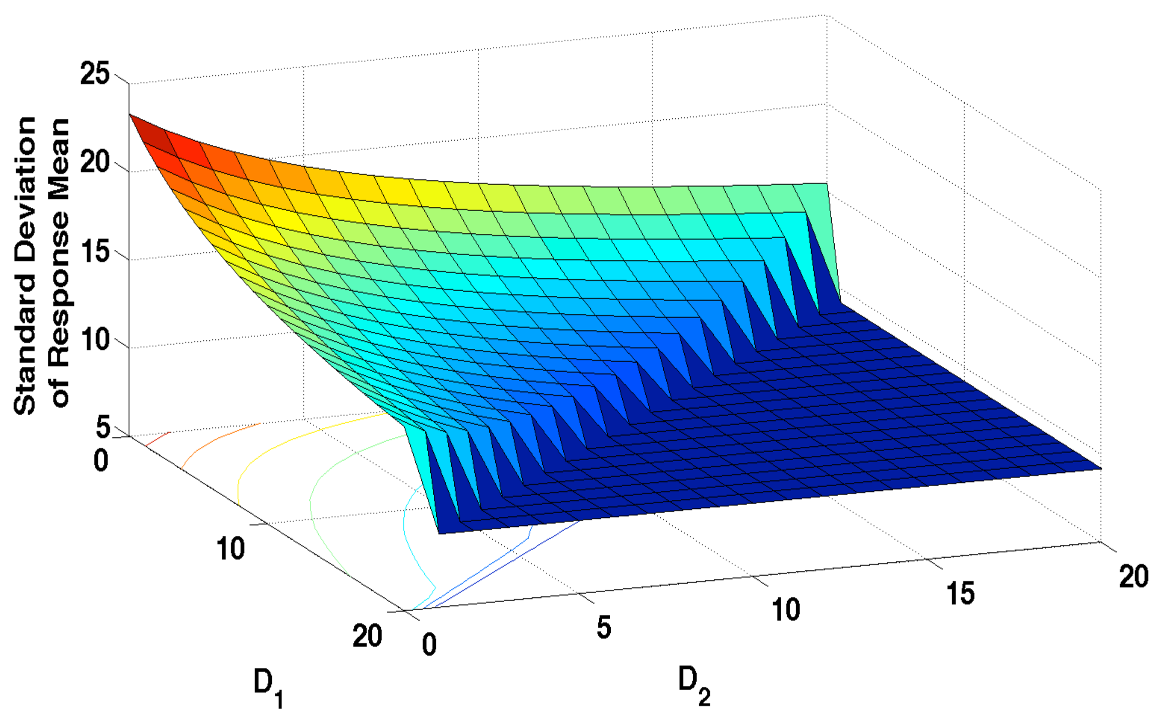

Figure 11. Design Space (Two variable - Case iv)

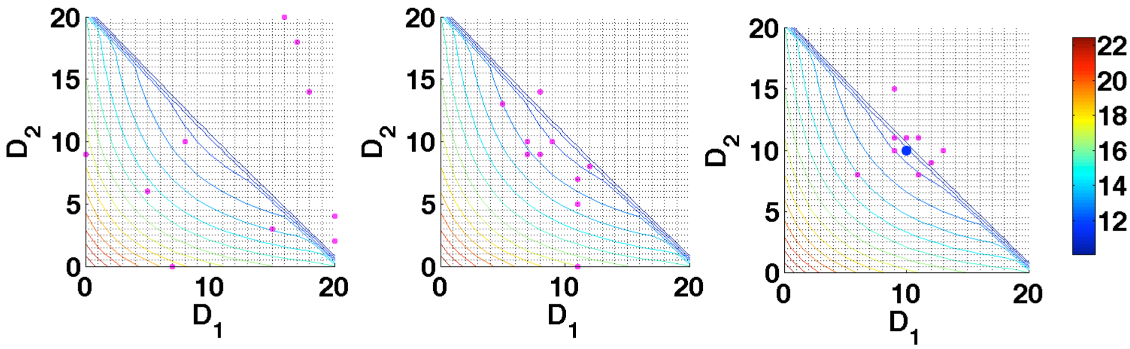

Figure 12. Particle movement (Case iv) a) Particles at first iteration, b) Particle at fifth iteration and c) Particles at the last iteration where the blue dot represents the optimum solution.

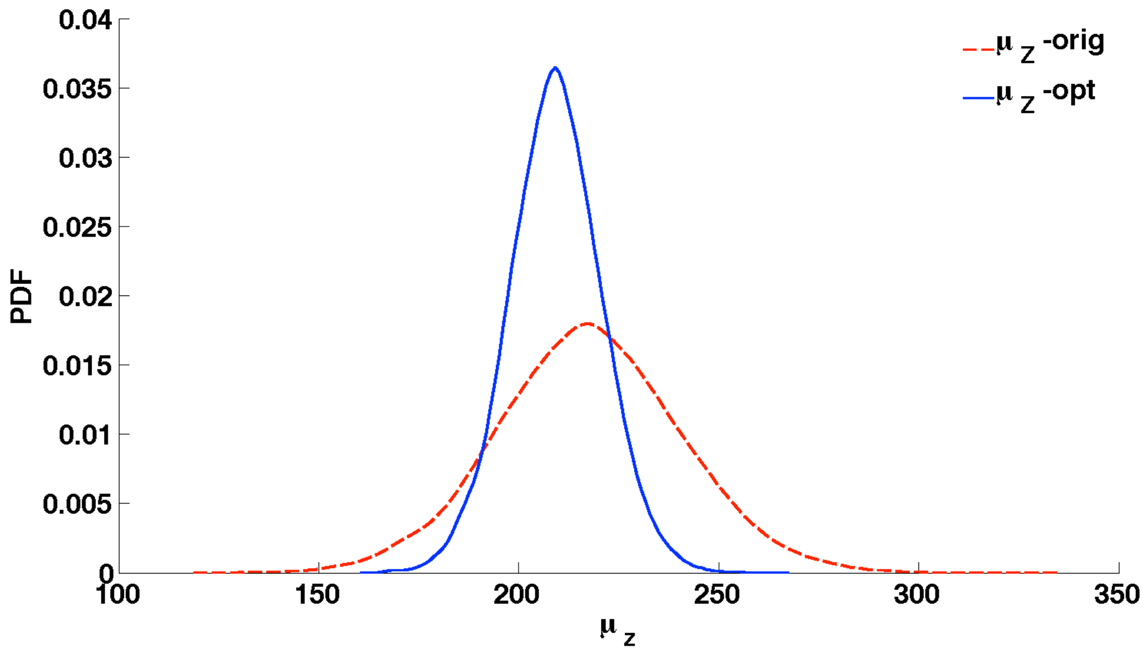

Figure 13. PDF of Response mean (Two variables – Case iv)

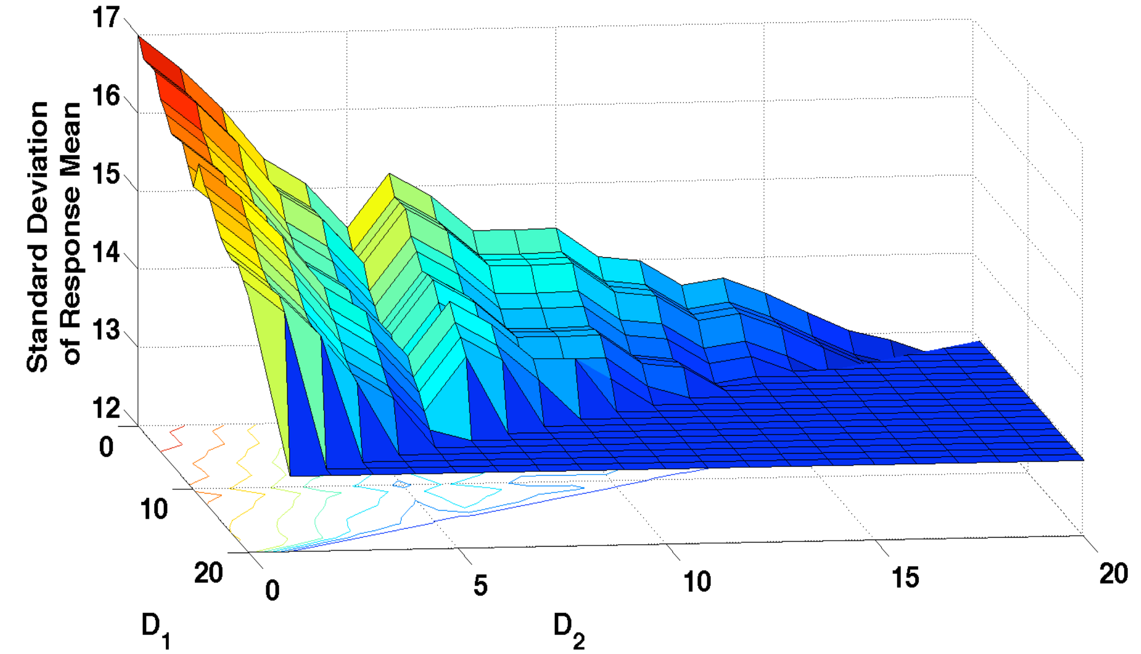

Figure 14. Design Space (Two variable - Case v)

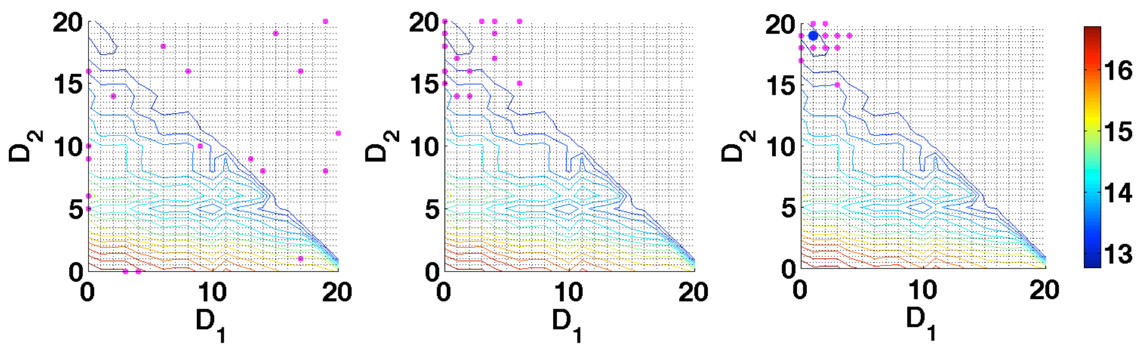

Figure 15. Particle movement (Case v) a) Particles at first iteration, b) Particle at fifth iteration and c) Particles at the last iteration where the blue dot represents the optimum solution.

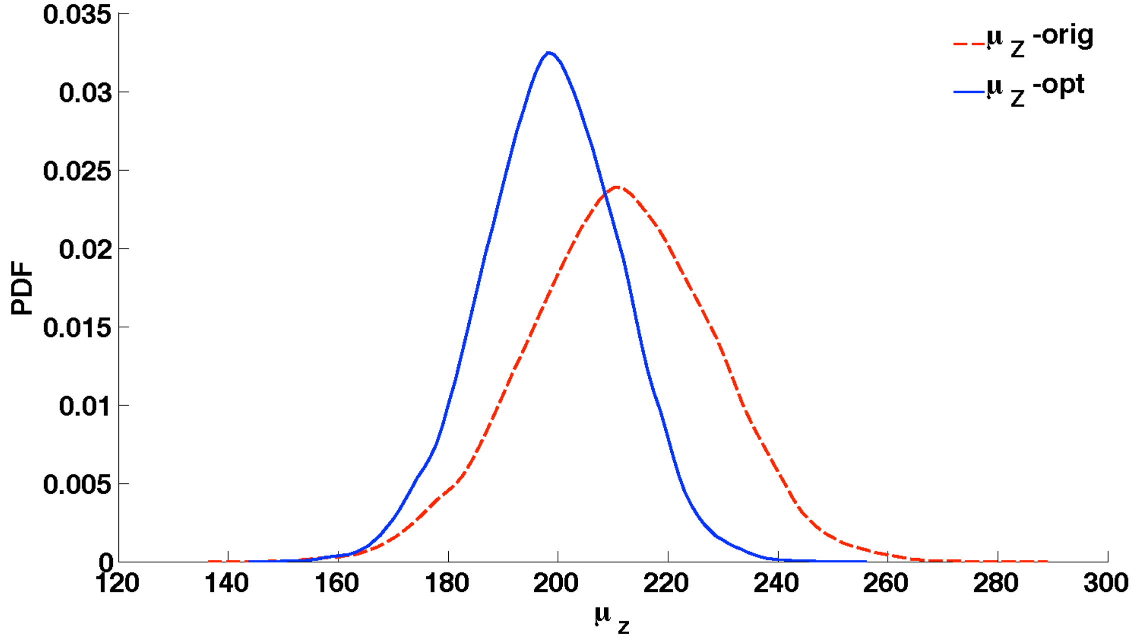

Figure 16. PDF of Response mean (Two variables – Case v)

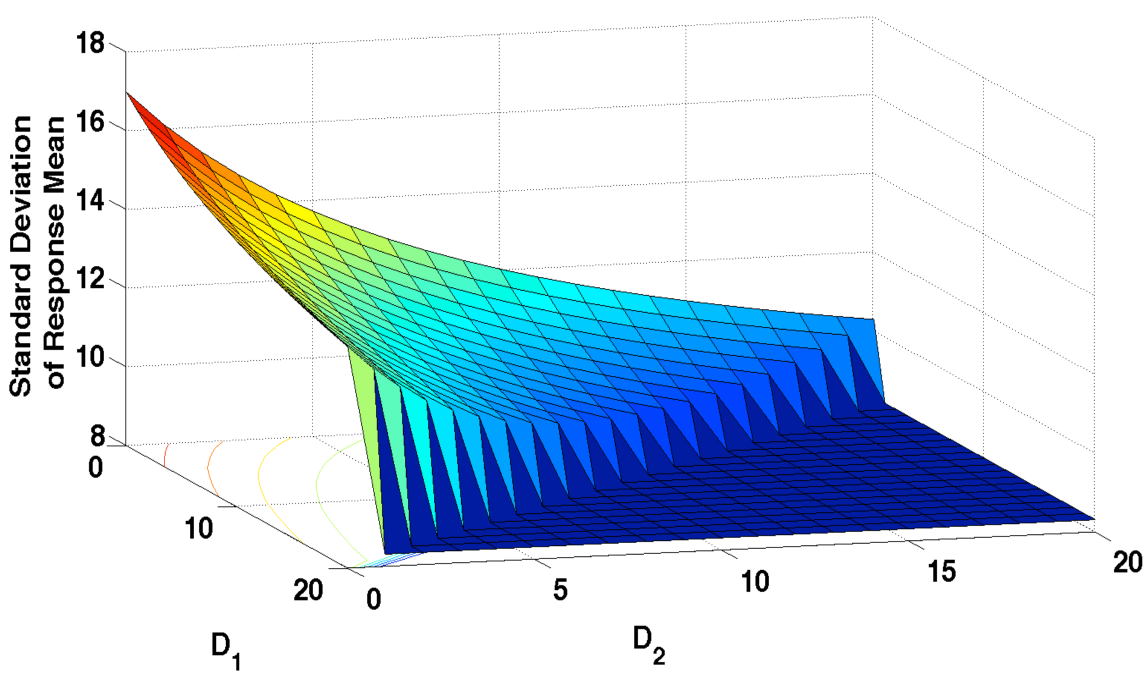

Figure 17. Design Space (Two variable - Case vi)

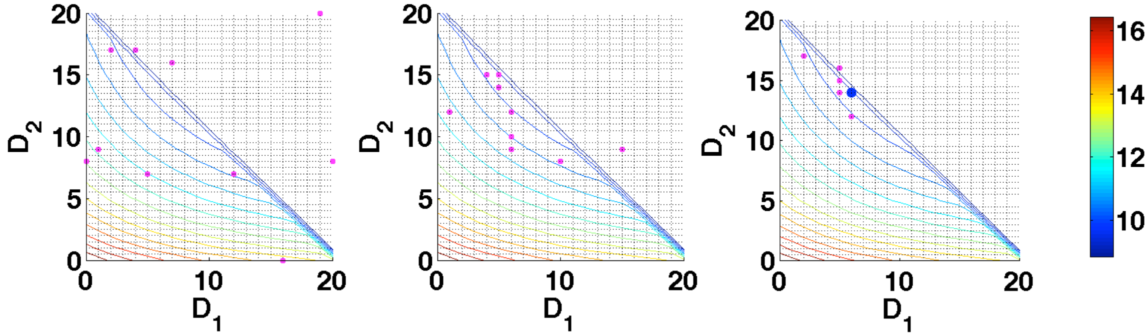

Figure 18. Particle movement (Case vi) a) Particles at first iteration, b) Particle at fifth iteration and c) Particles at the last iteration where the blue dot represents the optimum solution.

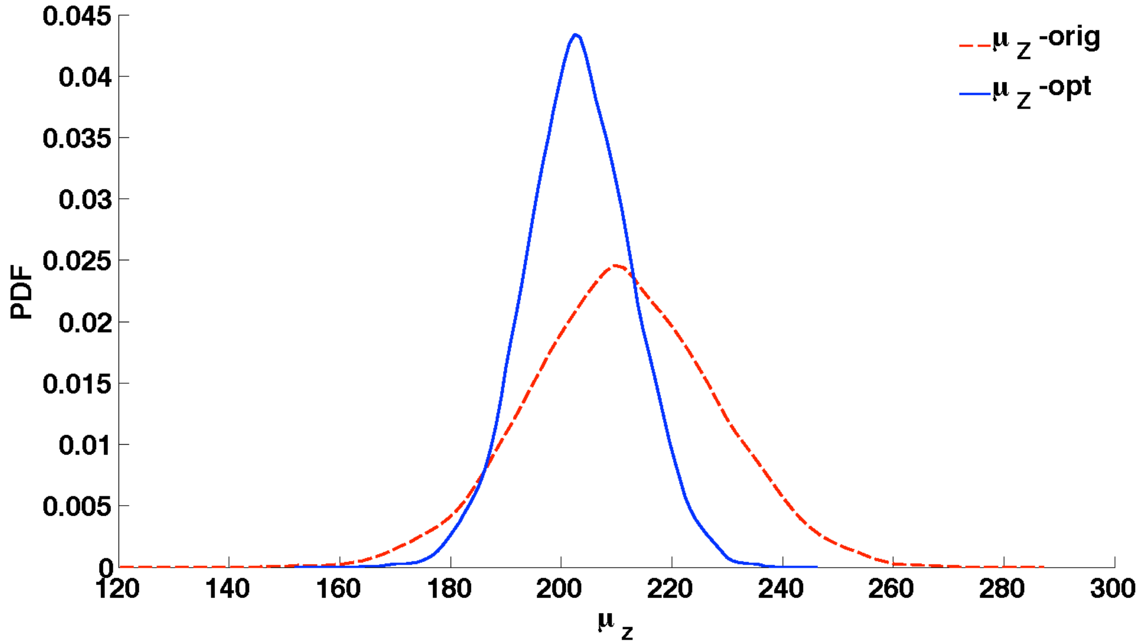

Figure 19. PDF of Response mean (Two variables – Case vi)

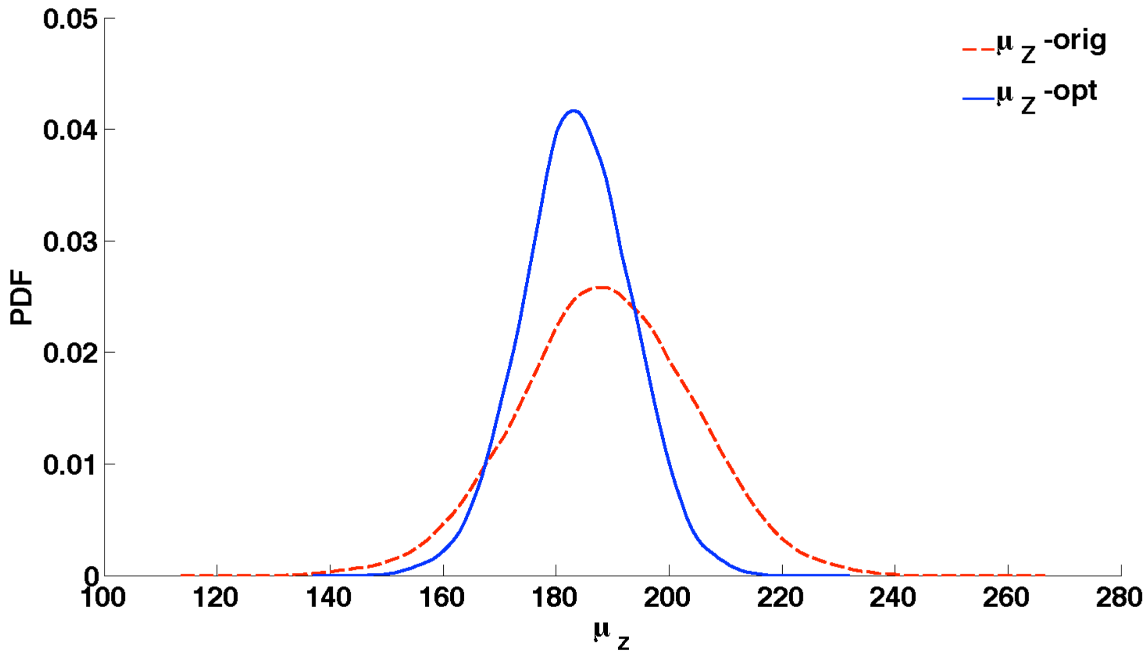

Figure 20. PDF of the Response mean (Three variables – Case vii)

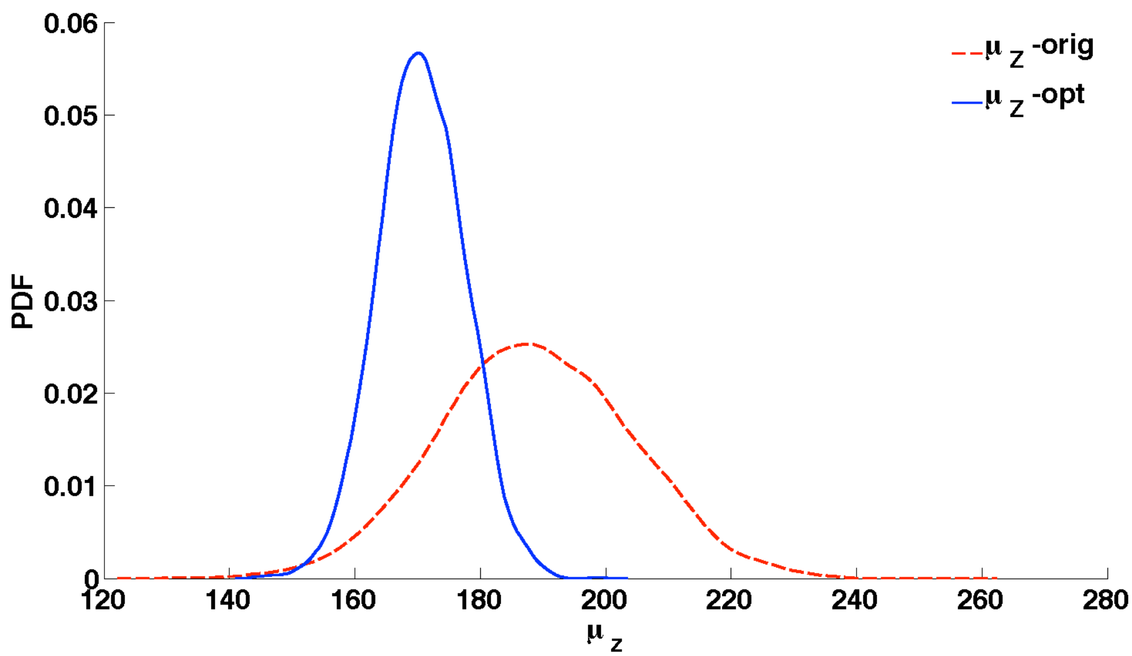

Figure 21. PDF of the Response mean (Three variables – Case viii)